EMMA TĂMÂIANU-MORITA,
SANDA CHERATA, CORNEL VÎLCU

# ANALIZA SINTAGMATICĂ A TEXTELOR ROMÂNEŞTI PRIN MIJLOACE INFORMATICE: PROIECTUL *SIASTRO*

## 1. CONSIDERAŢII INTRODUCTIVE

**1.1.** În societatea informatizată de astăzi, fiecare limbă are nevoie de produse tehnologice care să o conecteze la circuitul internaţional al tehnicilor şi instrumentelor de comunicare, stocare şi prelucrare a textelor în limbaj natural. Cristea, Tufiş 2002 subliniază cu claritate importanţa dezvoltării tehnologiilor limbajului uman pentru limba română:

> „Explozia de informaţii pe Internet, necesitatea tot mai mare de a accede la ele prin mijloace inteligente utilizând limba naturală proprie, indiferent de limba de depozitare a informaţiilor, face astăzi ca domeniul Prelucrării Limbajului Natural să fie considerat prioritar în Comunitatea Europeană. Tot mai multă lume recunoaşte că Tehnologia Limbajului va deveni domeniul primordial al Tehnologiei Informaţiei în secolul nostru. [...] Se manifestă, ca urmare, o necesitate stringentă de a ridica limba română, în privinţa resurselor şi a tipurilor de prelucrări automate, la nivelul altor limbi europene".

Tratarea cu mijloace automate a limbajului uman are, în prezent, caracteristici industriale, fiind încadrată în domeniul ***tehnologiile limbajului uman*** (**HLT,** *human language technology*). Resursele lingvistice pe care se bazează prelucrarea limbajului natural pot fi grupate în:

**(A)** *Resurse teoretice* (unele dintre cele mai importante fiind teoriile şi formalismele gramaticale);

**(B)** *Resurse de date lingvistice*:
- **(a)** *resurse textuale*: corpusuri; corpusuri adnotate (cele mai complexe fiind băncile de arbori – *tree banks*); corpusuri ale limbii vorbite;
- **(b)** *resurse lexicale*: dicţionare în format electronic; tezaure; glosare mono sau multilingve; concordanţe; terminologii; reţele semantice;
- **(c)** *resurse gramaticale*: descrieri ale gramaticii unei limbi utilizând un formalism gramatical.

**(C)** *Aplicaţii informatice*

*DACOROMANIA, serie nouă, XI–XII, 2006–2007, Cluj-Napoca, p. 77-87*

**(a)** aplicaţii de adnotare automată;
**(b)** aplicaţii de extragere de informaţii (*information extraction*) şi de regăsire a informaţiei (*information retrieval*);
**(c)** aplicaţii de asistare a elaborării documentelor (*authoring tools*): verificatoare ortografice şi gramaticale; verificatoare de limbă (*language checkers* sau *controlled language checkers*); aplicaţii de asistare orientate pe structuri (*structure-based authoring assistants*)
**(d)** aplicaţii de asistare a traducerii: traducere automată; traducere automată asistată de utilizator (*human-aided machine translation*); traducere automată statistică (*statistical machine translation*); traducere bazată pe exemple (*example-based translation*); aplicaţii de traducere umană asistată de calculator (printre cele mai importante, în această ultimă subcategorie, aflându-se memoriile de traducere).

Pentru ca resursele lingvistice existente să poată fi utilizate în aplicaţii noi aparţinând unor categorii diverse, trebuie definite practici şi formate comune de reprezentare, precum şi instrumente de conversie şi de adaptare a resurselor existente. În domeniul tehnologiei limbajului există **standarde** care reglementează această activitate, standarde pe care oricine dezvoltă o aplicaţie în acest câmp de lucru trebuie să le respecte, dacă doreşte să nu rămână izolat. Din păcate, aşa-numitele „limbi mici" sau „limbi mai puţin studiate", între care se numără şi limba română, au rămas insuficient frecventate şi reprezentate din acest punct de vedere. Faptul se explică, pe de o parte, prin aceea că mulţimea vorbitorilor este relativ restrânsă, iar resursele umane specializate în domeniu şi cele financiare sunt insuficiente, şi, pe de altă parte, prin particularităţile limbilor respective, care fac ca instrumentele deja existente să nu poată fi pur şi simplu preluate şi adaptate cu uşurinţă.

**1.2.** Pentru limbile de largă circulaţie se află în uz numeroase resurse din fiecare categorie menţionată anterior. În ceea ce priveşte limba română, resurse mai importante sunt :

**A. Resurse lingvistice**

**a.** ***Resurse lexicale***

a.1. *Dicţionare*

*Dicţionarul morfologic al limbii române,* elaborat de colectivul **RoLingva** de la S.C. Software ITC Cluj S.A., un dicţionar electronic care cuprinde descrierea lexico-morfologică şi fonologică a cuvintelor limbii române conţinute în DEX şi este acompaniat de un sistem de programe ce realizează analiza lexico-morfologică a cuvintelor româneşti, generarea paradigmei unei leme date, despărţirea în silabe etc.

În cadrul proiectului RORIC-LING a fost elaborat dicționarul morfologic de forme flexionare *RomDict*[1].

a.2. *Concordanțe*

Pot fi enumerate aici concordanța poeziilor lui B. Fundoianu, concordanța poeziilor lui George Bacovia[2], concordanța antumelor lui Mihai Eminescu.

a.3. *Terminologii*

Există atât terminologii – pe domenii de specialitate – pe suport de hârtie, cât şi baze de date terminologice. Sub egida Asociației Române de Terminologie, **TermRom**, s-a elaborat o bază de date terminologică multilingvă[3]. Institutul European din România, care a coordonat procesul de traducere a *acquis*-ului comunitar în limba română, deține o bază de date terminologică în continuă îmbogățire, consultabilă pe Internet la adresa http://www.ier.ro/. În cadrul masteratului de Terminologie-Traductologie al Catedrei de limbi moderne aplicate de la Facultatea de Litere a Universității „Babeş-Bolyai" s-a proiectat şi implementat o bază de date terminologică în domeniul legislației Uniunii Europene a Mediului (Cherata, Pop 2001a)[4]. De asemenea, există baze de date terminologice elaborate în cadrul ASE Bucureşti şi al Universității Bucureşti.

**b. *Resurse gramaticale***

Preocupări în domeniul lingvisticii informatice privitoare la limba română au apărut la începutul anilor 1990, în institute de cercetări (cum ar fi S.C. Software ITC Cluj S.A.) şi în marile centre universitare (Bucureşti, Iaşi, Cluj, Timişoara).

În prezent există atât lucrări teoretice care tratează modelarea unor aspecte ale limbii române în ***HPSG***, cât şi aplicații (analizoare morfologice operaționale, verificatoare ortografice încorporate în **Microsoft Office**, încercări de a realiza traducerea automată română-engleză şi engleză-română, programe de adnotare a textelor)[5]. Ca şi în alte subdomenii ale cercetărilor de lingvistică aplicată având drept obiect limba română, cercetările din domeniul tehnologiei limbii române s-au desfăşurat oarecum insular, fără ca echipele implicate să aibă cunoştință întotdeauna de demersurile întreprinse de ceilalți cercetători.

**B. Aplicații informatice**

Colectivul **RoLingva** de la S.C. Software ITC Cluj S.A. a elaborat un dicționar morfologic al limbii române şi un analizor morfologic complex[6]. La

[1] Vezi http://phobos.cs.unibuc.ro/roric/Ro/topic3.html

[2] Vezi Papahagi, Cherata, Tămâianu, Vuşcan 1999a şi 1999b.

[3] Vezi http://www.cimec.ro/PaginiGazduite/TR/default.htm

[4] Vezi şi Cherata, Pop 2001b, Cherata 2002.

[5] Contribuții valoroase au, de exemplu, Dan Tufiş, Dan Cristea, Emil Ionescu şi colaboratorii (vezi Tufiş 1999; Cristea, Tufiş 2002; Cristea, Ide, Romary 1998a şi 1998b; Ionescu 2001–2002a şi 2001–2002b; Cristea, Crăciun, Ursu 1998; Barbu, Ionescu 2001–2002; Barbu 2003).

[6] www.rolingva.ro. Vezi şi Şerban, Peev, Bibolar 1996, 1999 şi 2000; Peev, Bibolar, Jodal 1996.

Facultatea de Litere a Universității „Babeş-Bolyai" din Cluj-Napoca s-a implementat sistemul ***CONCORD*,** de realizare, cu mijloace electronice, a concordanțelor textelor poetice româneşti[7], cu ajutorul căruia s-au realizat concordanțele B. Fundoianu şi G. Bacovia.

În cadrul proiectului **RORIC-LING** s-au elaborat un model morfologic al limbii române, instrumente de adnotare a corpusurilor folosind gramaticile de dependențe, precum şi algoritmi de generare semiautomată a *synset*-urilor româneşti de tip ***WordNet***.

**1.3.** Pornind de la situația curentă în domeniu, Facultatea de Litere a Universității „Babeş-Bolyai" (UBB), S.C. Software ITC Cluj S.A. (SITC), Universitatea Tehnică din Cluj-Napoca (UTCN) şi Institutul de Lingvistică şi Istorie Literară „Sextil Puşcariu" (ILILC) au format un consorțiu care îşi propune să elaboreze un sistem informatic pentru analiza sintagmatică a textelor în limba română în cadrul proiectului **SIASTRO**[8]. Realizarea acestui obiectiv necesită cunoştințe din domenii diferite (lingvistică, lingvistică computațională şi grafică pentru realizarea interfețelor) şi presupune colaborarea unor echipe cu competențe interdisciplinare. Produsul informatic propriu-zis va fi însoțit de o bogată documentație de fundamentare teoretică şi descriere a sistemului sintagmelor limbii române.

Ne propunem să prezentăm în cele ce urmează coordonatele generale ale proiectului SIASTRO.

## 2. *SIASTRO*: OBIECTIVE PRINCIPALE ŞI COMPONENTE

**2.1.** Proiectul SIASTRO este configurat pe dimensiunea a trei obiective principale.

1. Realizarea unui ***sistem lexico-gramatical***, format din următoarele componente: (i) un *lexicon* cu intrări corespunzătoare cuvintelor limbii române, care conțin o serie de informații necesare tratării informatice a textelor, (ii) proceduri de analiză lexico-morfologică şi (iii) interfețele grafice corespunzătoare;
2. Modelarea analizei sintactice, pentru realizarea unui ***analizor sintagmatic***, care identifică şi analizează sintagmele nominale, verbale, adjectivale, adverbiale;
3. Realizarea unui sistem de *extragere a* ***termenilor***, ca primă aplicație practică a analizorului, compus, la rândul lui, din: (a) o componentă de identificare a sintagmelor care pot constitui termeni; (b) o bază de date terminologică în care se înscriu termenii validați de utilizator; (c) o componentă care realizează funcția de concordanță pentru sintagmele identificate, pentru a se stabili, pentru fiecare

[7] Vezi Cherata, Vuşcan, Tămâianu 1994a şi 1994b; Cherata 1996.

[8] Proiect finanțat prin grantul CEEX nr. 86-CEEX-03-II/31.07.2006 (august 2006 – august 2008). Director de proiect: conf. dr. Emma Tămâianu-Morita; responsabili echipe de cercetare: CSI Luciana Peev, conf. dr. ing. Rodica Potolea, CSI dr. Felicia Şerban.

candidat de termen, contextele în care acesta apare; (d) o interfață care să permită utilizatorului comunicarea cu sistemul.

**2.2.** Se valorifică prin aceasta experiența specifică în domeniu a fiecărui partener. După cum menționam anterior, colectivul din cadrul UBB a realizat implementarea, cu mijloace automate, a concordanțelor textelor poetice românești, cu contexte pentru cuvinte-lemă (nu și pentru sintagme). În cadrul SITC s-a realizat un dicționar morfologic al limbii române (care nu conține însă atributele necesare analizei sintagmatice) și un analizor lexico-morfologic, fundamentarea teoretică fiind asigurată de ILILC. În cadrul UTCN s-au realizat interfețe grafice și *site*-uri web la un înalt nivel profesional. Numai corelarea acestor componente permite realizarea unui sistem informatic complex de tipul celui vizat în proiectul de față.

Proiectul SIASTRO se va finaliza prin realizarea unui prototip de extractor de termeni, însoțit de o documentație comprehensivă privind (i) aspectele formale ale gramaticii limbii române și (ii) modalități de implementare a acestora în crearea de produse informatice.

În termeni concreți, proiectul va contribui la îmbogățirea resurselor informatice din perimetrul limbii române, și anume (a) *resursele lexicale* – prin îmbogățirea *Dicționarului morfologic al limbii române* realizat de SITC cu atributele necesare analizei sintactice; (b) *resursele gramaticale* – prin descrierea formală a sintagmelor limbii române și implementarea unui analizor sintactic; (c) *aplicațiile informatice* – prin crearea unui sistem de extragere a termenilor și a contextelor lor dintr-un corpus dat.

Prin respectarea normelor și standardelor în reprezentarea uniformă a informațiilor, rezultatele proiectului devin compatibilizabile cu sisteme lexicale și terminologice multilingve dezvoltate pe scena internațională a cercetării aplicate din domeniul tratării automate a limbajului natural.

**2.3.** Noutatea pe care o aduce proiectul SIASTRO constă în aceea că, pentru limba română, pune pentru prima dată problema integrării, într-o aplicație informatică, a prelucrării limbajului natural și a terminologiei, efectuând în acest fel un prim pas spre integrarea prelucrării limbii române în aplicațiile terminologice. Am ales ca primă aplicație tocmai *extragerea termenilor,* deoarece aceasta vine în sprijinul creării de terminologii – un domeniu de stringentă actualitate pentru limba română.

Pe plan internațional se acordă un interes tot mai mare unor aplicații de acest tip. Sisteme precum **XPLANATION, Xerox Terminology Suite, TRADOS** conțin instrumente de extragere a termenilor. Dacă până nu de mult cercetările din domeniul prelucrării limbajului natural și cele din domeniul terminologiei și terminografiei se desfășurau independent unele de altele, la ora actuală se caută mijloace și metode pentru a atinge dezideratul de integrare a resurselor lingvistice. Astfel, în aplicațiile de tratare a limbajului natural sunt prevăzute mecanisme de tratare a termenilor, iar în aplicațiile terminologice sunt integrate instrumente de

tratare a limbajului natural. În ultimii ani se fac eforturi considerabile pentru reprezentarea uniformă a resurselor. Consorțiul **SALT** (**S**tandards-based **A**ccess service to multilingual **L**exicons and **T**erminologies[9]), format din grupuri academice, guvernamentale, comerciale şi asociații din Europa şi SUA a fost creat tocmai pentru aceasta şi a elaborat formatul **XLT** (***XML representation of Lexicons and Terminologies***). În acelaşi scop se desfăşoară activitatea grupului **OSCAR** (**O**pen **S**tandards for **C**ontainer/Content **A**llowing **R**e-use") din cadrul asociației **LISA** (**L**ocalisation **I**ndustry **S**tandards **A**ssociation[10]). În momentul de față se vorbeşte despre *gestiunea resurselor lingvistice* (*language resource management*), iar sub auspiciile ISO (Organizația Internațională de Standardizare) s-a înființat comitetul **TC 37/SC4**, care are drept sarcină elaborarea de standarde în acest domeniu. În documentul cu titlul *Dealing with language matters – Possible interactions between ISO TC37 and ISO JTC1/SC36*[11] al comitetului respectiv sunt reprezentate sursele de informații lingvistice şi interacțiunea dintre ele.

În realizarea proiectului SIASTRO ne propunem să ne aliniem standardelor în vigoare pentru reprezentarea uniformă a informațiilor, astfel încât limba română să poată fi integrată în sistemele multilingve existente.

### 3. Metodologie şi rezultate

**3.1.** Fiecare dintre cele trei obiective menționate sub **2.1.** are, în fapt, finalitate practică, permițând deschideri ulterioare înspre alte aplicații. Dintre cele trei componente ale **SIASTRO**, central este analizorul sintagmatic: în funcție de cerințele realizării lui se proiectează lexiconul şi se implementează analiza lexico-morfologică, iar rezultatele furnizate de el sunt preluate de componenta de extragere terminologică. La rândul lui, analizorul sintagmatic îşi defineşte obiectivele în funcție de cerințele extragerii terminologice: tipurile de sintagme analizate sunt cele care constituie structura posibililor termeni.

În primul rând, pentru realizarea sistemului lexico-gramatical este necesară dezvoltarea unui *lexicon* cu atribute sintactice. Lexiconul va fi creat plecând de la *Dicționarul morfologic al limbii române*, elaborat de SITC, unde, pentru fiecare intrare, se vor adăuga seturi ierarhice de atribute necesare analizei sintactice. Pe baza lor se vor implementa procedurile de analiză lexico-morfologică care să furnizeze structurile informaționale pentru analizorul sintactic. În etapa I a proiectului, desfăşurată în anul 2006, echipele de lingvişti şi informaticieni au operat o amplă analiză comparativă a celor mai utilizate formalisme de descriere a structurilor sintagmatice: **HPSG**, **LFG**, gramatici de dependențe, gramatici categoriale etc., precum şi a posibilităților de reprezentare a informațiilor lexicale, ținând seama de standardele existente pe plan internațional: **TEI, OLIF, XLT** etc. (vezi lista bibliografică). Seturile de atribute corespunzătoare fiecărei clase lexico-

[9] http://www.ttt.org/salt/
[10] http://www.lisa.org/
[11] http://www.tc37sc4.org/new_doc/ISO_TC_37-4_N051_Dealing_with_language_matters.pdf

gramaticale vor fi reprezentate în lexicon în conformitate cu formalismul ales. Acest prim obiectiv se va finaliza, aşadar, cu un lexicon suficient de bogat pentru a permite extragerea terminologică.

Modelarea analizei sintagmatice se fundamentează, de asemenea, pe examinarea comparativă a modelelor de analiză sintactică, pentru a alege formalismul adecvat cerinţelor specifice ale limbii române. Definirea formalismului impune stabilirea tipurilor şi subtipurilor de sintagme, din perspectiva relevanţei lor pentru identificarea termenilor, şi descrierea detaliată a structurii fiecăreia. Astfel, pentru fiecare tip şi subtip de sintagmă se va descrie structura arborelui sintactic, decorat cu atribute lexico-gramaticale: trăsăturile centrului sintagmei, valenţele unităţilor componente, trăsături locale etc. Aceste elemente vor servi la elaborarea algoritmilor de realizare a analizei sintagmatice. Implementarea analizorului sintagmatic se va finaliza, aşadar, cu o aplicaţie de analiză sintactică pentru sintagmele de categorii majore (sintagmele nominale, adjectivale, verbale şi adverbiale), prevăzută cu o interfaţă care să vizualizeze arborii sintactici obţinuţi.

Obiectivul final, implementarea extragerii termenilor, se va finaliza cu o aplicaţie informatică de sine stătătoare, care identifică termenii-candidaţi dintr-un corpus de texte de specialitate şi permite utilizatorului inserarea lor în baza de date, împreună cu informaţiile terminologice privind termenul, partea de vorbire şi contextele de apariţie, însoţite de sursele respectivelor contexte.

Sistemul de extragere de termeni proiectat este un sistem interactiv. El realizează identificarea candidaţilor de termeni din corpusul care i se prezintă spre analiză, urmând ca utilizatorul să stabilească, pe baza contextelor de utilizare, care dintre candidaţi se confirmă drept termeni. Fiecare termen validat de utilizator va fi înscris, în formă canonică, într-o bază de date terminologică, unde utilizatorului i se va da posibilitatea de a adăuga şi/sau modifica atributele relevante (clasă lexico-gramaticală, alte atribute gramaticale, registru de utilizare, statut etc.). La cererea utilizatorului, contextele de utilizare pe care le selectează vor fi şi ele inserate în baza de date, împreună cu specificarea surselor de apariţie. Structura bazei de date terminologice va fi proiectată în conformitate cu standardele ISO.

**4. Dezvoltări posibile, aplicaţii, utilizări**

Tehnologia realizată în cadrul proiectului **SIASTRO** va genera rezultate atât în domeniul teoretic (contribuţii la realizarea sistemelor-expert de analiză lingvistică), cât şi în domeniul practic (îmbogăţirea resurselor pentru limba română: lexicale, gramaticale şi aplicaţii informatice). Cercetările vor putea fi continuate, fără îndoială, înspre analiza sintactico-semantică a textelor româneşti, cu aplicaţii dintre cele mai diverse: corectoare gramaticale, sisteme de asistare a învăţării limbii române (atât de către vorbitorii nativi, cât şi de către cei străini), sisteme de adnotare a corpusurilor, sisteme de traducere automată etc.

Proiectul este conceput în aşa fel încât rezultatele sale să poată fi valorificate, într-o primă etapă, în multiple domenii ale activităţii specifice din mediul educaţional şi de cercetare, legate de prelucrarea textelor româneşti. Beneficiile

posibile privesc asistarea învăţării limbii române şi creşterea gradului de corectitudine în utilizarea limbii, mai ales în perimetrul textelor scrise.

În plus, sistemul lexico-gramatical, analizorul sintagmatic şi sistemul de extragere a termenilor constituie baza unor aplicaţii ulterioare care se adresează şi unei game mai largi de utilizatori, inclusiv din mediul economic, prin crearea de instrumente care să faciliteze procesarea şi verificarea textelor redactate sau traduse în limba română, mai cu seamă a textelor specializate. Astfel, prin componenta sa terminologică, proiectul de faţă vine în sprijinul utilizatorilor care se confruntă în mod curent, prin natura activităţii lor, cu un volum mare de texte specializate, întrucât le furnizează o primă organizare a corpusului respectiv pe domenii şi subdomenii. Prin aceasta se reduce, implicit, volumul de timp şi efort pe care utilizatorul ar fi nevoit să îl investească în respectiva activitate.

Posibili beneficiari ai rezultatelor concrete ale proiectului şi ai dezvoltărilor lui ulterioare se găsesc în rândul persoanelor fizice sau juridice care desfăşoară activităţi terminologice, de documentare, de traducere a textelor specializate şi de interpretariat, al persoanelor interesate în a învăţa limba română (mai ales pentru documentaţii scrise), în didactica limbii române, al celor ce lucrează în meseriile cărţii, al tuturor celor care prelucrează corpusuri mari de texte.

Un grupaj de studii aflat în pregătire pentru numărul următor al revistei „Dacoromania” va propune o primă ilustrare a câtorva teme din perimetrul activităţilor de documentare şi efectuare a descrierii lingvistice, precum şi din perimetrul celor de formalizare a informaţiei, desfăşurate de membrii echipelor participante la proiectul **SIASTRO**. Corelate, articolele vor oferi cititorului o imagine de ansamblu asupra dimensiunilor în care proiectul va genera nu numai rezultatele aplicative programate, ci şi contribuţii relevante din punct de vedere teoretic-descriptiv, în direcţia unei explicări funcţionale coerente a sistemului structurilor sintagmatice din limba română, concretizată într-un model utilizabil şi în dezvoltări aplicative ulterioare.

## BIBLIOGRAFIE

Barbu 2003 = Ana-Maria Barbu, *Sintaxa determinatorilor. Descriere HPSG asistată de calculator*, Bucureşti, http://www.racai.ro/~abarbu/carte

Barbu, Ionescu 2001–2002 = Ana-Maria Barbu, Emil Ionescu, *Teorii Gramaticale Contemporane: Gramatica Centrilor De Sintagma (HPSG)*, http://phobos.cs.unibuc.ro/roric/papers/hpsgro.doc

Cherata 1996 = Sanda Cherata, *Sistem de realizare a concordanţelor limbii poetice româneşti*, în D. Tufiş (ed.) *Limbaj şi tehnologie*, Editura Academiei Române, 1996, p. 215-220.

Cherata 2002 = Sanda Cherata, *TeRo – a terminological database for a multilingual terminology on the environment*, în Rodica Baconsky, Daniel Gouadec, Gheorghe Lascu (eds.), *Teritorii actuale ale traducerii – Teritoires actuels de la traduction, Actele Colocviului internaţional – Actes de Colloque International – Traduire l'Europe, Universitatea „Babeş-Bolyai”, Cluj-Napoca, 9–10 martie 2001*, Editura Echinox, Cluj-Napoca.

Cherata, Pop 2001a = Sanda Cherata, Liana Pop (coord.), *Glosar poliglot al legislaţiei mediului: engleză – franceză – germană – română*, Cluj-Napoca, Editura Echinox.

Cherata, Pop 2001b = Sanda Cherata, Liana Pop (coord.), *Bază de date terminologică elaborată la Universitatea „Babeş-Bolyai" din Cluj-Napoca / Base de données terminologiques élaborée à l'Université „Babeş-Bolyai" de Cluj-Napoca***,** în „T&T – Terminologie et traduction. La revue des services linguistiques des institutions européennes", 2.

Cherata, Vuşcan, Tămâianu 1994a = Sanda Cherata, Teodor Vuşcan, Emma Tămâianu, SILEX – *Un sistem lexico-morfologic computerizat pentru analiza textelor româneşti*, în DR, serie nouă, 1994–1995, nr. 1–2, p. 54-66.

Cherata, Vuşcan, Tămâianu 1994b = Sanda Cherata, Teodor Vuşcan, Emma Tămâianu, *SILEX – Funcţiile de lematizare şi de generare a paradigmelor*, în DR, serie nouă, 1994-1995, nr. 1–2, p. 67-80.

Cristea, Crăciun, Ursu 1998 = Dan Cristea, Ovidiu Crăciun, Cristian Ursu, *GLOSS – A Visual Interactive Tool for Discourse Annotation*, în *Proceedings of the Workshop on Annotation Tools, ESSLLI'98*, Saarbruecken, August 1998, http://folli.loria.fr/cds/1998/pdf/krenn/corpus-workshop/gloss.pdf

Cristea, Ide, Romary 1998a = Dan Cristea, Nancy Ide, Laurent Romary, *Marking-up Multiple Views of a Text: Discourse and reference,* în *Proceedings of the First International Conference on Language Resources and Evaluation,* Granada, May 1998, www.cs.vassar.edu/~ide/papers/discourse.granada.pdf

Cristea, Ide, Romary 1998b = Dan Cristea, Nancy Ide, Laurent Romary, *Veins Theory: A Model of Global Discourse Cohesion and Coherence*, în Christian Boitet, Pete Whitelock (eds.), *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, San Francisco, California, Morgan Kaufmann Publishers, p. 281-285, citeseer.ist.psu.edu/cristea98veins.html

Cristea, Tufiş 2002 = Dan Cristea, Dan Tufiş, *Resurse lingvistice româneşti şi tehnologii informatice aplicate limbii române*, în O. Ichim, F.-T. Olariu (eds.), *Identitatea limbii şi literaturii române în perspectiva globalizării*, Romanian Academy, Institute of Romanian Philology „A. Philippide", Iaşi, Trinitas Publishing House, http://thor.info.uaic.ro/~dcristea/papers/cristeatufis2002.zip

Ionescu 2001–2002a = Emil Ionescu, *Head-Driven Phrase Structure Grammar – A General Presentation*, phobos.cs.unibuc.ro/roric/papers/hpsg.doc

Ionescu 2001–2002a = Emil Ionescu, ***Premise** ale unui dicţionar morfologic electronic al limbii române*, http://phobos.cs.unibuc.ro/roric/papers/rofullformro.doc

Papahagi, Cherata, Tămâianu, Vuşcan 1999a = M. Papahagi, S. Cherata, E. Tămâianu, T. Vuşcan, *Concordanţa poeziilor lui B. Fundoianu*, Cluj-Napoca, Editura Echinox.

Papahagi, Cherata, Tămâianu, Vuşcan 1999b = M. Papahagi, S. Cherata, E. Tămâianu, T. Vuşcan, *Concordanţa poeziilor lui George Bacovia*, Cluj-Napoca, Editura Echinox.

Peev, Bibolar, Jodal 1996 = Luciana Peev**,** Lidia Bibolar, Endre Jodal, *Model de formalizare a morfologiei limbii române*, în vol. *Limbaj şi tehnologie,* Bucureşti, p.67-72.

Şerban, Peev, Bibolar 1996 = Felicia Şerban**,** Luciana Peev**,** Lidia Bibolar, *Bază de date a limbii române. Fonetică şi fonologie*, în *Limbaj şi tehnologie*, Bucureşti, p. 157-160.

Şerban, Peev, Bibolar 1999 = Felicia Şerban**,** Luciana Peev**,** Lidia Bibolar, *La base de données de la langue roumaine*, în Terminometro: *La terminologie en Roumanie et en République de Moldova,* hors-série, nr. 4, Cluj-Napoca, Editura Clusium, 2000, p. 40-42.

Şerban, Peev, Bibolar 2000 = Felicia Şerban**,** Luciana Peev**,** Lidia Bibolar, *Bază de date a limbii române*, în *Terminologia în România şi Republica Moldova*, Cluj-Napoca, Editura Clusium, p. 37-38.

Tufiş 1999 = Dan Tufiş, *Yet Another Head Driven Generator of Natural Language*, în *International Journal on Information and Control*, vol. 3, ICI, Bucharest, http: //www.ici.ro/ici/revista/sic1999_3/art03.html

**Aplicaţii de extragere a termenilor**

**ProMemoria (BridgeTerm)**: http://www.bridgeterm.com/en/promem.html (memorie de traducere cu componentă de extragere de termeni).

**Xerox TermFinder**: (Xerox Multilingual Knowledge Management Solutions) Part of Xerox Terminology **Suite (XTS)** http://www.mkms.xerox.com/

**TRADOS**: www.**trados**.com (memorie de traducere, componente de recunoaştere terminologică şi de extragere de termeni).

**Standarde de reprezentare a informației**

**OLIF** – **O**pen **L**exicon **I**nterchange **F**ormat, elaborat de OLIF Consortium, http://www.olif.net/, un format care se concentrează asupra interschimbului de date între resursele lexicografice ale diferitelor sisteme de traducere automată.

**TEI** – **T**ext **E**ncoding **I**nitiative, http://www.tei-c.org/P4X/index.html

**TEI-Term** – **T**erminology **I**nterchange **F**ormat, elaborat de **TEI** (**T**ext **E**ncoding **I**nitiative), http://www.tei-c.org/Vault/GL/P3/TE.htm

**XLT** – elaborat de **S**tandards-based **A**ccess to multilingual NLP **L**exicon and human-oriented **T**erminology resources (**SALT)**, http://www.loria.fr/projets/SALT/. SALT este un consorțiu EU/US care a dezvoltat modelul XLT/DXLT, un format de interschimb între datele lexicale utilizate de aplicațiile NLP şi datele terminologice şi care ulterior a evoluat în formatul **TBX**.

**TMF** – **T**erminology **M**arkup **F**ramework, ISO 16642, un metamodel care defineşte structurile şi mecanismele necesare reprezentării informatice a datelor terminologice, proiectat pentru a exprima schimbul de informații între diferite TML-uri (Terminology Markup Languages). Conține, ca TML-uri, standardele GENETER şi MSC (varianta MARTIF conformă cu TMF).

## AUTOMATIC PHRASE ANALYSIS FOR ROMANIAN TEXTS: THE *SIASTRO* PROJECT
(*Abstract*)

In the present context of the information society, each language needs technological products to connect it to the international environment of computerized communication tools, and natural-language text processing and storage techniques. The linguistic resources necessary for human language technologies (HLT) applications are classified into three main types: (a) theoretical resources (such as grammatical theories and formalisms); (b) linguistic data resources (textual, lexical and grammatical resources); (c) computer applications (such as automatic annotation, information extraction and information retrieval applications, authoring tools, translation authoring assistants). The article starts from an analysis of the current situation in the field regarding the existing resources for the Romanian language, and proceeds to the presentation of a complex interdisciplinary project (*SIASTRO*) undertaken by a consortium of four partners from Cluj-Napoca, aiming at the creation of a system for the automatic phrase analysis of Romanian texts. The system is designed to have three components: (1) *a lexico-grammatical system*, which consists of a lexicon with entries corresponding to Romanian words and containing sets of data required for the automatic processing of texts, lexico-morphological analysis procedures and the necessary graphic interfaces; (2) *a parser*, which performs the analysis of noun phrases, verb phrases, adjectival phrases, adverbial phrases; (3) an interactive *system for term extraction* from specialized texts, as a first practical application of the parser. The

project's expected outcome is the implementation of a prototype system for term extraction, as well as comprehensive scientific documentation concerning both the formal aspects of Romanian grammar, and modalities of implementation. Based on these results, research can subsequently be extended towards the syntactic-semantic analysis of Romanian texts, with most diverse applications: grammar checking programs, systems for computer-assisted Romanian language learning, both for native, and for non-native speakers, corpus annotation systems.

*Universitatea „Babeş-Bolyai"*
*Facultatea de Litere*
*Cluj-Napoca, str. Horea, 31*